Statistical Visualization of Environmental Data on the Web Using nViZn

Lacey Jones

Michigan State University
Dept. of Statistics and Probability
A505 Wells Hall
East Lansing, MI 48823
e-mail: joneslac@msu.edu

Jürgen Symanzik

Utah State University
Dept. of Mathematics and Statistics
3900 Old Main Hill
Logan, UT 84322-3900
e-mail: symanzik@sunfs.math.usu.edu

Abstract

Statistical analyses of large-scale data can often be hard to interpret. Often several pages of numbers are used to describe the data, which makes it tedious or difficult to find the numerical output that is of interest. Converting these numbers into information that is understandable and useful to someone without an extensive statistical background is also a task that is not easily accomplished. Visual representations such as maps, graphs, and charts can aid in this process. These help to better understand the information and processes the data is explaining. A recent improvement in visual statistics has been the use of the Internet. The new Illumitek software tool, nViZn, the follow-up to the Graphics Production Library (GPL), is a tool that allows a programmer to create interactive statistical displays on the Web. These displays allow a user to expand or narrow the focus of a graphical representation. It is possible to order, overlay, and rearrange the format of the data as fast as the user's Internet connection can process. Users do not need a statistical computing background to work with nViZn displays on the Web. The Internet provides easy access to such statistical displays to anyone with an Internet connection and a graphical user interface. We tested nViZn in order to determine its ability to display quality, information-rich statistical graphics. We were also interested in the difficulty of the process needed to create these displays. Our main application aimed at the display of 148 hazardous air pollutants (HAPs) for the 60,803 census tracts in the continental United States obtained from the U.S. Environmental Protection Agency's (EPA) Cumulative Exposure Project (CEP).

Keywords

Graphics Production Library, JAVA, Statistical Graphics, Statistical Software, WWW.

1. Introduction

Large-scale data sets often contain a great amount of valuable information. However, this information is frequently hidden within an overwhelming amount of numbers. A statistical analysis of large-scale data sets often involves many complicated procedures and numerical summaries. Putting the results of these procedures and numerical summaries into a form that is easy to understand and useful to people of various backgrounds can be difficult. Large-scale data such as remote sensing data or transactional data (e.g., purchases in stores or phone calls made) is becoming more

widely collected with improvements in access speed and memory space for databases. Hence, our ability to interpret and relay information about such data sets must also improve. Statistical visualizations such as charts, plots, data linked maps, and tables are necessary tools in this process of relaying information. They can successfully describe the data in a way that informs a knowledgeable statistician. In addition, statistical visualizations can often enlighten those without extensive knowledge on the techniques used in statistical analyses. By using nViZn (Wilkinson, Rope, Carr and Rubin 2000), the latest software for statistical displays on the Web, combined with the technology of the Internet, we become able to create interactivity. We can expand or narrow the focus of a graphical representation. We can order, overlay, and rearrange the format of the data. These features are available to anyone with an Internet connection.

In Section 2 of this paper, we provide a general overview on nViZn, a new software product for the visualization of data on the Web. In Section 3, we describe how we might use nViZn for the U.S. Environmental Protection Agency's (EPA) hazardous air pollutants (HAP) data. We finish with a discussion in Section 4.

2. Visualization and Interactivity Through nViZn

The new software product by Illumitek (Website: http://www.illumitek.com), nViZn, is a JAVA-based software development kit (SDK), related to the book "The Grammar of Graphics" by Leland Wilkinson (1999). nViZn (Wilkinson, Rope, Carr and Rubin 2000) is a follow-up to an older statistical visualization tool developed within the Bureau of Labor Statistics (BLS) called the Graphics Production Library (GPL) (Carr, Valliant and Rope 1996). The GPL was initially created to allow the BLS to present statistical summaries over the Web through interactive charts and tables. nViZn has all the original capabilities of the GPL. One of its additional capabilities allows a programmer to create data-linked micromaps, i.e., a link of row labeled univariate or multivariate statistical summaries to corresponding geographical regions (Carr and Pierson 1996, Carr, Olsen, Courbois, Pierson and Carr 1998, Carr, Olsen, Pierson and Courbois 2000). nViZn has been designed to handle large data sets of nearly any format and the user can complete analytics within the SDK. It is also possible to represent a visual object in 3D or add animation to show changes over time in the data.

The graphics in Figures 1 and 2 demonstrate a small proportion of the capabilities nViZn has to offer. The data sets used to create these graphics have been taken from the examples of the statistical graphics program XGobi (Swayne, Cook and Buja 1998) (Website: http://www.research.att.com/areas/stat/xgobi).

Figure 1 shows six nViZn graphs - a boxplot, two histograms, two scatter plots, and a line plot - all created from data sets obtained from the XGobi Website. The first row of graphs was created from a data set entitled Flea and consists of a boxplot of head sizes for different flea species and a histogram of head sizes for all species. The second row of graphs was created from a data set entitled Olive and contains a scatter plot of linolenic and eicosanoic acid amounts in olives and a line plot of linoleic acid amounts for different types of olives. The final row of graphs was created from a data set entitled Places and shows a scatter plot of education levels and art levels for places within the US and a histogram of education level. The user can set colors and formats for each of these graphs. Included in the format options of nViZn are a 3D appearance, animation, and popup windows to reveal data values for specified variables when pointed to by the computer mouse. Interactivity also exists to rearrange the variables in the graphics. The user can add or remove variables. Brushing techniques allow to link graphics in such a way that when a user highlights a set of values in one graphic the values are also highlighted in all linked graphics.

Figure 2 is a paneled scatter plot in nViZn. This scatter plot was created with the Flea data obtained from the XGobi Website and plots species by aedeagus width for fleas. Paneled scatter plots have all of the capabilities listed above for the graphs in Figure 1.

nViZn is a JAVA-based software development kit. Once the coding has been written for an application, end users of this application can interactively extract the desired information through an Internet browser without purchasing the nViZn software or gaining the know-how to create the graphics themselves. As for most programming languages, much of the nViZn coding for application developments is reusable among common applications. Often, only minor changes are required.

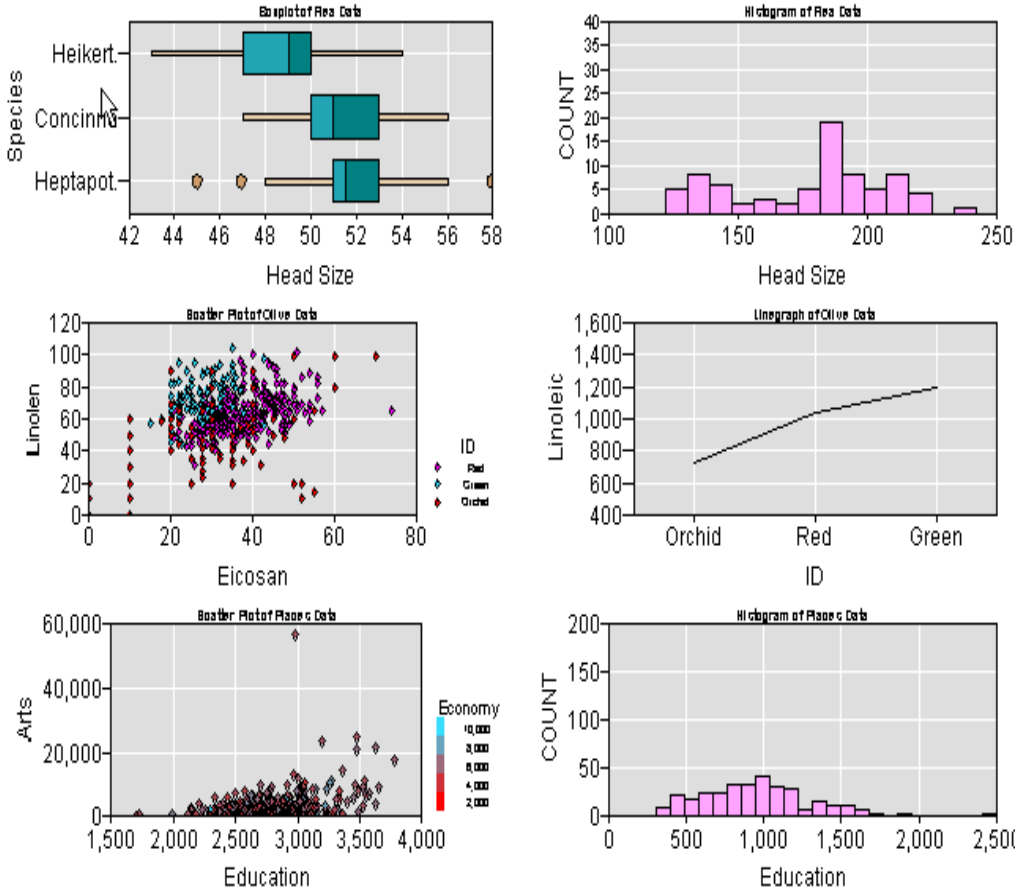

Figure 1: Samples of boxplots, histograms, scatter plots,
and line plots using three different data sets from the XGobi Website.

Figure 3 is an example of nViZn coding. This code produces the graphic in Figure 2. Lines 1-8 list the files of nViZn that are needed to make this graphic. Choice of graph formats, interactivity, data types, and analytics determine whether additional files are needed. Lines 9-18 and 28-30 set up the application and the frame to display the graphic. Lines 19-27 and 31-38 deal with the type of data used and how it is read into the application. Since nViZn allows for most types of data files, these lines will differ for each type of data file. Lines 39-47 specify the type of graphic and the variables used in the graphic. The remaining lines create the popup windows that reveal data values of specified variables when the user moves the computer mouse over any of the displayed points.

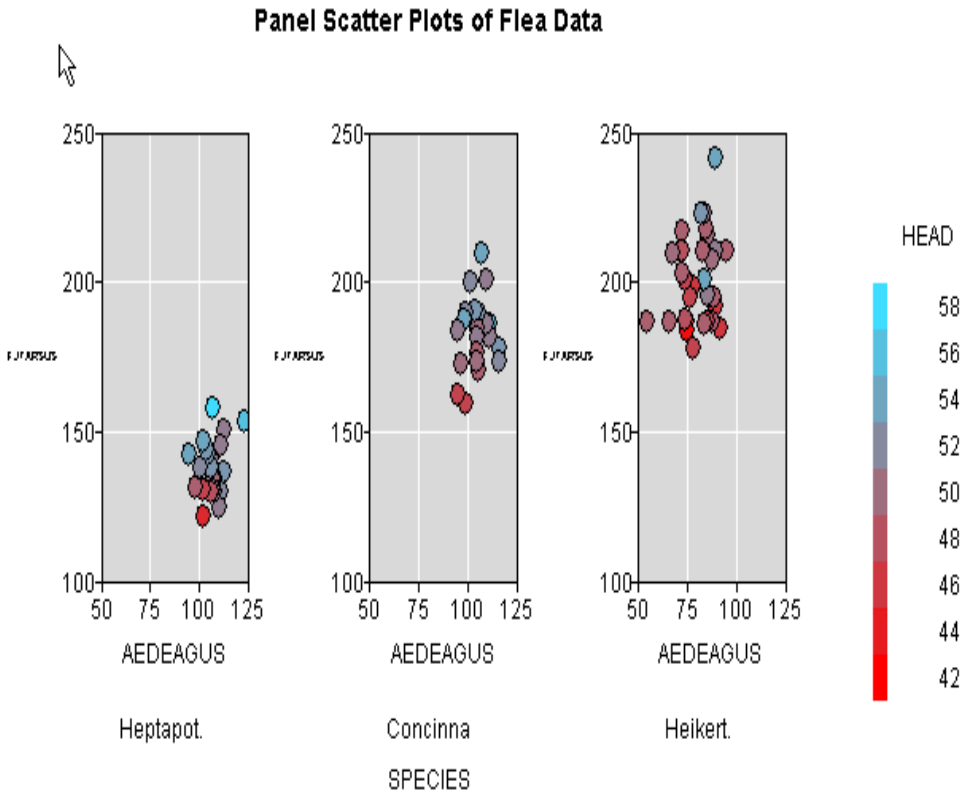

Figure 2: Paneled scatter plot of flea species (from the XGobi Website) comparing sizes of body parts.

```
1   import com.illumitek.gpl.dataview.primitive.*;
2   import com.illumitek.gpl.dataview.*;
3   import com.illumitek.gpl.specification.*;
4   import com.illumitek.gpl.graph.*;
5   import com.illumitek.gpl.controllers.meta.*;
6   import com.illumitek.gpl.elements.*;
7   import java.awt.*;
8   import java.rmi.*;

9   public class FleaPanel{

10  private DataView dataview;
11  private GPLGraph graph;

12  public static void main(String[] argv) {

13          new FleaPanel();
14  }

15  public FleaPanel() {

16      Frame frame = new Frame();
17      frame.setSize(600,450);
18      frame.setLayout(new BorderLayout());

19      try {
20        rebuildDataView();
21        rebuildGraph();
22        wireListeners();

23        dataview.beginDataPass();
24      }
25      catch (RemoteException ex) {
26        ex.printStackTrace();
27      }

28      frame.add("Center", graph);
29      frame.show();
30    }

31    public void rebuildDataView() throws RemoteException {

32      if (dataview == null) {
33        dataview = new FlatFileDataView();
34      }

35      FlatFileSourceSpecification source = new
36        FlatFileSourceSpecification("flea");
37      dataview.setSource(source);
38    }

39    public void rebuildGraph() {

40      if (graph == null)
41        graph = new GPLGraph();

42      graph.setExpression("AEDEAGUS*FJTARSUS*SPECIES");
43      graph.setTitle("Panel Scatter Plots of Flea Data");

44      Points point = new Points();
45      point.setColor("HEAD");
46      graph.addElement(point);
47    }

48    public void wireListeners() throws RemoteException {

49      dataview.addDataPassListener(graph);

50      MetaDataPopupDisplayController metaPopUp =new
51      MetaDataPopupDisplayController();
52      graph.addPropertyChangeListener(metaPopUp);
53    }
54  }
```

Figure 3: nViZn code used to produce the graphic in Figure 2.

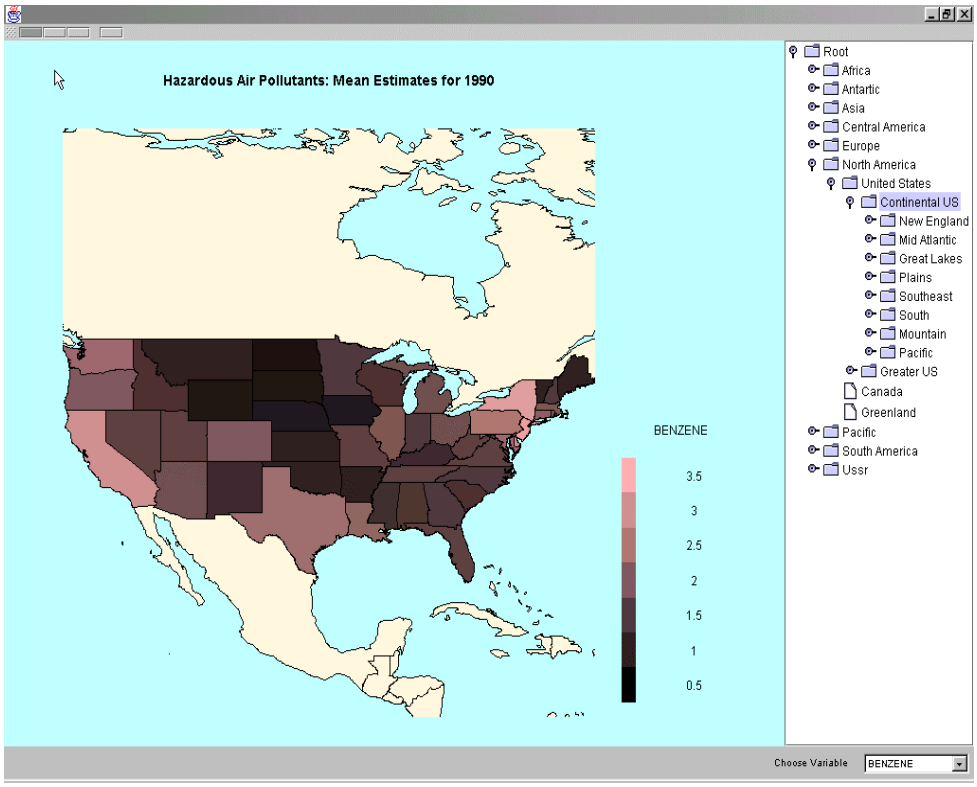

Figure 4: Drill-down map of the U.S. for any HAP.

3. Application to EPA's HAP Data

The main data set used in this paper to demonstrate the capabilities of nViZn has been obtained through the U.S. Environmental Protection Agency's (EPA) Cumulative Exposure Project (CEP) (Website: http://www.epa.gov/CumulativeExposure). The Cumulative Exposure Project has been conducted to determine the amount of toxins Americans were exposed to through air, food, and water intake. The air toxic section of the data contains modeled concentration estimates with uncertainty bounds of 148 hazardous air pollutants (HAPs) for the 60,803 census tracts in the continental United States. These estimates are annual averages for the year 1990, and are described in Rosenbaum, Axelrad, Woodruff, Wei, Ligocki, and Cohen (1999). The underlying statistical analysis is explained in Woodruff, Axelrad, Caldwell, Morello-Frosch and Rosenbaum (1998) and Caldwell, Woodruff, Morello-Frosch and Axelrad (1998). The EPA had originally planned to publish this data set on the Web in December 1998. Work was done on a text-based Web page with graphics and interactions using the Graphics Production Library (GPL) (Carr, Valliant and Rope 1996). However, the EPA abandoned its intention to publish this data on an interactive Web site due to the outdating of the data in 1999 (Symanzik, Carr, Axelrad, Wang, Wong and Woodruff 1999). Further details on the EPA CEP Website have been described in Symanzik, Axelrad, Carr, Wang, Wong and Woodruff (1999) and Symanzik, Wong, Wang, Carr, Woodruff and Axelrad (2000).

For EPA's HAP data, a hierarchy of interactive data-linked micromaps would be a good approach to visualize this complex data set. A user could start with the entire

country and drill down to the county and census tract level. Through the use of interactive software, the data could be presented in such a way that finding the estimates of a region of interest and comparing any particular HAP for different regions would be easy and concise. An interactive graphical user interface that can be created with nViZn gives the option to link these micromaps with line plots, scatter plots, and histograms. Hence, such an application would allow the easy comparison of the estimates between all census tracts of a county and the underlying distribution for a chosen HAP. Along with the comparison of any particular HAP, the ability to compare many or all HAPs for a chosen census tract also could be implemented.

Figure 4 shows a drill-down map of the U.S. with the average modeled 1990 benzene concentration. In this interactive application which is based on an original nViZn example, a user could drill-down to the state and county levels using either the menu on the side panel or double clicking on the designated area. The menu box on the bottom allows users to select the HAP of their choice. Alterations may be made to the color and formatting of the map.

Figure 5 is an example of a data-linked micromap. The modeled 1990 benzene concentration estimates are plotted for each of the counties in Pennsylvania. Application interactions for this micromap might include popup data windows, pan and zoom, and resorting of the data to allow for easy comparison by county. As mentioned before, color and formatting options are available.

Figure 6 shows a table produced with nViZn, showing the mean modeled 1990 benzene concentration for the upper half of the counties (from Figure 5) within Pennsylvania. The table alternates in color from row to row, which makes matching the table values to their respective row and column categories fast and easy. Alterations may be made to tables that customize color schemes and table formats to match the nature of the data. In a future application, the tables may also be made interactive allowing the user to add and remove variables or sort and rearrange their order to facilitate an easy comparison of the data.

4. Discussion

From working with this software, it is apparent that nViZn offers many capabilities that allow one to display and visually analyze statistical data that may originate from government, academia, and business. However, it was quite difficult to learn how to develop applications of statistical visualizations that use nViZn. The main reason for these difficulties is the current deficit in documentation on the methods and syntax of the SDK. The nViZn developers certainly need to take major steps towards making the package "developer friendly". Otherwise, widespread use of the product is unlikely. Undoubtedly, as more people use nViZn, the difficult aspects of its use should easily be recognized, which will provide directions towards its improvement.

One of the problems that arose while trying to re-create example applications and develop our own applications was telling the computer where to look for files, i.e., how to set classpaths. No part of the nViZn documentation provided the proper procedure for doing this. It was necessary to search for several methods in JAVA tutorial books and the online JAVA documentation. After a considerable amount of reading and much experimentation, the classpaths were set correctly and the examples could be run.

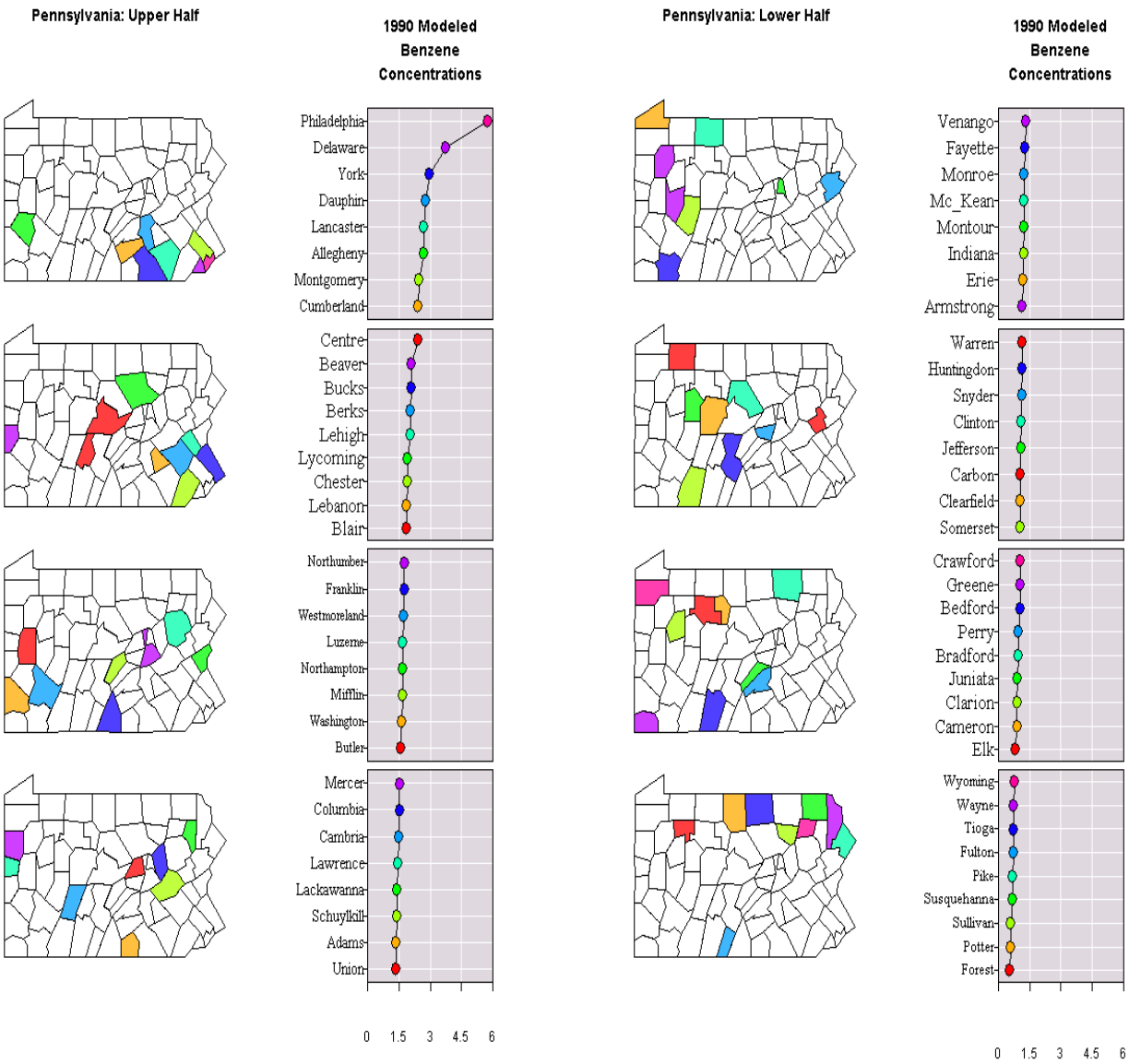

Figure 5: Data-linked micromap of modeled 1990 benzene concentrations of Pennsylvania by county.

Pennsylvania: 1990 Modeled Benzene Concentrations
Upper Half

	Minimum	Mean	Maximum	Quartile1	Median	Quartile3
Philadelphia	2.79	5.75	15.9	4.19	5.16	6.44
Delaware	1.98	3.76	18.91	2.69	3.21	4.13
York	0.94	2.96	12.07	1.23	1.72	3.9
Dauphin	0.92	2.79	6.17	1.54	2.63	3.61
Lancaster	0.91	2.68	10.38	1.26	1.92	3.39
Allegheny	1.18	2.68	8.74	2.01	2.58	3.07
Montgomery	1.24	2.45	4.24	2.0	2.43	2.86
Cumberland	0.87	2.42	7.4	1.12	2.33	2.74
Centre	0.64	2.4	9.82	0.72	1.31	3.56
Beaver	1.0	2.09	4.63	1.19	1.81	2.79
Bucks	1.14	2.08	14.34	1.74	2.01	2.22
Berks	0.88	2.07	5.71	1.22	1.96	2.73
Lehigh	0.82	2.03	4.4	1.42	1.94	2.51
Lycoming	0.6	1.93	5.92	0.75	1.03	2.91
Chester	0.96	1.91	5.66	1.34	1.77	2.11
Lebanon	1.01	1.86	3.4	1.21	1.44	2.46
Blair	0.72	1.86	4.39	0.97	1.81	2.45
Northumber	0.79	1.75	3.02	0.88	1.82	2.79
Franklin	0.76	1.75	3.71	0.93	1.06	3.12
Westmoreland	0.72	1.73	4.87	1.08	1.49	2.21
Luzerne	0.7	1.69	4.06	0.98	1.59	2.31
Northampton	0.94	1.68	2.9	1.3	1.65	1.98
Mifflin	0.75	1.67	3.96	0.81	0.9	2.78
Washington	0.74	1.63	4.47	1.01	1.2	2.29
Butler	0.71	1.57	4.8	0.96	1.09	1.83
Mercer	0.68	1.52	2.55	0.87	1.35	2.17
Columbia	0.7	1.52	3.64	0.8	0.96	2.12
Cambria	0.69	1.51	4.07	0.79	1.14	2.13
Lawrence	0.83	1.46	2.71	0.96	1.3	1.86
Lackawanna	0.77	1.42	2.08	1.14	1.44	1.68
Schuylkill	0.76	1.4	3.36	0.87	1.04	1.79
Adams	0.8	1.36	3.95	0.95	1.01	1.43
Union	0.7	1.35	3.94	0.8	0.84	1.26
Venango	0.61	1.31	3.86	0.69	0.86	1.57

Figure 6: Table of top mean modeled 1990 benzene concentration estimates for Pennsylvania.

Another problem was to obtain the proper plug-ins to run interactive visualizations. Only the most recent plug-in was compatible for running nViZn. Help was also required from one of the developers of nViZn to figure out how to format data in such a way that it could be read in and used by the application. Other difficulties arose in translating sample nViZn code, such as for making a micromap, to match the needs of our application. Some documentation existed for most of the nViZn sample code. However, this documentation contained very little details and was often unclear or misleading. Without further documentation, learning how to use the SDK by trial and error alone is slow and difficult. Other than the need for more documentation on nViZn, it is also necessary that software developers who want to develop applications with nViZn have a solid background in programming with JAVA.

Obviously, our implementation of the EPA HAP application is only a prototype and not a final product. However, based on our experiences gained during this implementation we can conclude that nViZn has much potential for interactive, Web-based statistical displays of data from government, academia, and business. However, it is difficult to learn. To facilitate the learning of nViZn, Illumitek has begun to offer training courses and has recently released a new and improved version of nViZn with added documentation. For continued product improvement the company is gaining feedback from people testing the product, such as with this project, and from its users. Whether nViZn becomes widely used or not, it will pave the way for ongoing progress in the field of interactive statistical visualization on the Web.

Acknowledgments

Lacey Jones' work was supported in part by an U.R.C.O. Grant from the Office of the Vice President for Research from Utah State University. The work of Jürgen Symanzik was supported in part by the NSF "Digital Government" (NSF 99-103) grant #EIA-9983461 and by a New Faculty Research Grant from the Vice President for Research Office from Utah State University.

References

Caldwell, J. C., Woodruff, T. J., Morello-Frosch, R. & Axelrad, D. A. (1998), 'Application of Health Information to Hazardous Air Pollutants Modeled in EPA's Cumulative Exposure Project', *Toxicology and Industrial Health* 14(3), 429-454.

Carr, D. B., Olsen, A. R., Courbois, J. P., Pierson, S. M. & Carr, D. A. (1998), 'Linked Micromap Plots: Named and Described', *Statistical Computing and Statistical Graphics Newsletter* 9(1), 24-32.

Carr, D. B., Olsen, A. R., Pierson, S. M. & Courbois J. P. (2000), 'Using Linked Micromap Plots to Characterize Omernik Ecoregions', *Data Mining and Knowledge Discovery* 4(1), 43-67.

Carr, D. B. & Pierson, S. (1996), 'Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps', *Statistical Computing and Statistical Graphics Newsletter* 7(3), 16-23.

Carr, D. B., Valliant, R. & Rope, D. (1996), 'Plot Interpretation and Information Webs: A Time-Series Example from the Bureau of Labor Statistics', *Statistical Computing and Statistical Graphics Newsletter* 7(2), 19-26.

Rosenbaum, A. S., Axelrad, D. A., Woodruff, T. J., Wei, Y. H., Ligocki, M. P. & Cohen, J. P. (1999), 'National Estimates of Outdoor Air Toxics Concentrations', *Journal of the Air and Waste Management Association* 49, 1138-1152.

Swayne, D. F., Cook, D. & Buja, A. (1998), 'XGobi: Interactive Dynamic Data Visualization in the X Window System', *Journal of Computational and Graphical Statistics* 7(1), 113-130.

Symanzik, J., Axelrad, D. A., Carr, D. B., Wang, J., Wong, D. & Woodruff, T. J. (1999), 'HAPs, Micromaps and GPL - Visualization of Geographically Referenced Statistical Summaries on the World Wide Web', *in* 'Annual Proceedings (ACSM-WFPS-PLSO-LSAW 1999 Conference CD)', American Congress on Surveying and Mapping.

Symanzik, J., Carr, D. B., Axelrad, D. A., Wang, J., Wong, D. & Woodruff, T. J. (1999), 'Interactive Tables and Maps - A Glance at EPA's Cumulative Exposure Project Web Page', *1999 Proceedings of the Section on Statistical Graphics*, American Statistical Association, Alexandria, VA, 94-99.

Symanzik, J., Wong, D., Wang, J., Carr, D. B., Woodruff, T. J. & Axelrad, D. A. (2000), 'Web-Based Access and Visualization of Hazardous Air Pollutants', *Geographic Information Systems in Public Health: Proceedings of the Third National Conference August 18-20, 1998, San Diego, California*, Agency for Toxic Substances and Disease Registry, http://www.atsdr.cdc.gov/GIS/conference98/.

Wilkinson, L. (1999), *The Grammar of Graphics*, Springer-Verlag, New York, NY.

Wilkinson, L., Rope, D. J., Carr, D. B. & Rubin, M. A. (2000), 'The Language of Graphics', *Journal of Computational and Graphical Statistics* 9(3), 530-543.

Woodruff, T. J., Axelrad, D. A., Caldwell, J. C., Morello-Frosch, R. & Rosenbaum, A. S. (1998), 'Public Health Implications of 1990 Air Toxics Concentrations Across the United States', *Environmental Health Perspectives* 106(5), 245-251.